第 6 章

实验研究方法

张岩　徐飞　奚恺元

学习目标

1. 学习实验设计的基本原则
2. 了解实验设计的评判标准
3. 掌握典型的实验设计

6.1 研究的类型

科学研究林林总总，但是总是会涉及理论和数据。大体上，根据理论和数据的关系，可以把研究归为三类：第一类是有数据支持但无理论指导（data without theory）的研究，第二类是有理论指导但无数据支持（theory without data）的研究，第三类是既有理论指导又有数据支持（theory with data）的研究。

首先来看第一类有数据支持但无理论指导的研究。例如，你通过调查发现，中国人喜欢吃米饭，美国人喜欢吃土豆；中国人喜欢喝茶，美国人喜欢喝咖啡；中国人喜欢吃豆沙包，美国人喜欢吃奶酪蛋糕。尽管有这些发现，但是并没有一个理论能够帮助你解释为什么中国人和美国人在饮食上存在这样的差异。此外，这些发现也不能帮助你预测中国人和美国人对于其他饮食的偏好，当然也不能预测其他国家的人对饮食的偏好。这样的研究就属于有数据支持但无理论指导的研究。

再看第二类——有理论指导但无数据支持的研究。假如你有一个理论，描述一个人领到的奖金和他的工作效率之间的关系。根据这个理论，你建立了一个模型。你的模型有很多非常漂亮的参数，能够把奖金和工作效率的关系完全量化。看上去你似乎可以精确地预测多少奖金可以带来多高的工作效率。但是问题在于，你并没有实证的数据来检验自己的理论和模型到底对不对，也就是说，你没有办法知道你的预测在现实生活中到底成不成立。这样的研究就属于有理论指导但无数据支持的研究。需要指出的是，在这里我们所说的数据是实证数据，是从现实中得来的，而不是根据你的模型计算出来的数据。

令人惋惜的是，很多管理学和经济学的研究往往落入以上两类。第一类的文章往往有满页的表格、整段的事实，但研究本身仅仅停留在数据层面，而没有上升为理论。这个问题在很多管理学的研究中较为普遍。第二类的文章恰恰相反，整篇只有理论建模而没有实证，例如很多经济学的文章。这两类研究类型在科学研究中都是不可取的。

可取的研究应该既有理论的指导，又有数据的支持，即我们开篇所说的第三类研究。例如，我们的理论认为人们在预测别人偏好的时候，往往会将自己的偏好强加于别人。由这个

理论我们可以导出很多预测，例如“中国人因为自己喜欢吃中国菜，因而更容易高估喜欢吃中国菜的美国人的比例”。为了验证这个预测是不是正确，我们让中国人首先回答他们自己是不是喜欢吃中国菜，然后让他们估计喜欢吃中国菜的美国人的比例；同时，让美国人回答他们是不是喜欢吃中国菜，从而得出喜欢吃中国菜的美国人的真实比例。这样我们就可以获得一套数据并用统计方法来检验理论和数据是不是相符合。同理，假定你根据你的理论建立了一个关于奖金和工作效率的模型。为了检验你的模型是否正确，首先你需要找到一群人，给他们不同数额的奖金，并观察他们的工作效率。其次，你把奖金的数额放到你的模型里面去，你的模型就会预测出人们的工作效率水平。最后，把模型预测的工作效率水平和上面调查得到的真实工作效率水平相对比，就可以发现你的模型是否符合实际。这样，我们就完成了一个既有理论指导、又有数据支持的研究。一个研究只有同时拥有理论指导和数据支持，才可能经得起检验。

6.2　理论和假设

那么，到底什么是理论呢？理论就是解释和预测某些现象的一系列假设（Schweigert，2006），通常被用来解释已经发生的事件及预测未来的事件。在科学研究中，我们需要用数据来支持待验证的理论，或者用理论来解释现有的数据。

假设是关于自变量（independent variable）和因变量（dependent variable）之间关系的陈述，用以解释某个现象。这里所说的现象就是因变量，而导致这个现象的因素是自变量。例如，你的假设是“使用大的电脑显示器能够提高员工的工作积极性”，那么员工的工作积极性就是因变量，而电脑显示器的大小则是自变量。这个例子将贯穿这一部分接下来的内容，我们用它来解释与假设相关的一些概念。

6.2.1　自变量

什么是自变量呢？自变量就是在你的假设中引起某个现象的变量，也是实验中可以被实验者控制的变量。在“显示器与工作积极性”的例子中，显示器的大小就是自变量。实验者通过改变显示器的大小，来检验显示器的大小是否会影响员工的工作积极性。

显示器会有不同的尺寸，同样，自变量通常会拥有几个不同的取值，每一个取值就叫作自变量的一个水平（level）。自变量的取值可以分为有限的和无限的，也可以分为离散的和连续的。有的自变量有有限个离散的取值。显示器的大小就是这样，我们现在在市面上能买到的显示器只有有限的几个尺寸，并且它的大小也不可能是连续变化的。而有的自变量则可以是连续的。比方说工作时间就是一个连续的变量。一般在实验中，我们并不能检测一个连续变量的所有可能的值，而是会选取其中的部分值来检验自变量对因变量的影响。如果你的假设是“随着工作时间的增加，员工的工作效率会降低”，那么通常情况下我们会选取 2—3 个工作时间点，比如 1 个小时、4 个小时、7 个小时，将它们作为自变量的三个取值。

6.2.2　因变量

因变量就是在你的假设中被预测的变量，或者实验者认为会随着自变量变化而变化的变量。在“显示器与工作积极性”的例子中，员工的工作积极性就是因变量。

如果对自变量和因变量之间关系的描述要上升到理论阶段，我们通常认为自变量和因变量之间存在因果关系（causality）。比方说，你假设“朋友多的人比朋友少的人更幸福”。这个假设仅仅是一个相关性（correlation）的假设，说的是朋友多少和幸福水平高低的关系。但是这个假设没有说明朋友多少和幸福高低是否存在因果关系。是因为有更多朋友，人们更幸福呢，还是因为人们更幸福，所以更可能交到朋友呢？一个相关性的假设是无法回答这个问题的。相比之下，如果一个假设说“增加朋友的数量会提高人们的幸福水平”，这就是一个因果关系的描述，朋友数量是自变量，幸福水平是因变量。

理论的一个重要特征就是它的假设描述了变量之间的因果关系，而不仅仅是相关关系。因果关系对于理论的建立是非常重要的。拿上面的例子来说，弄清楚是不是朋友的数量影响了幸福水平，或者，可以让我们知道什么因素可以影响人们的幸福水平，从而更深入地研究为什么这些因素影响人们的幸福水平。同时，弄清因果关系也可以帮助我们对如何提高幸福水平提出实质性的建议。

6.2.3　几种简单假设的形式

从自变量数量的角度来看，最简单的假设是单一自变量假设。在单一自变量假设中，最为简单的情况是这个自变量只有两个取值。比方说，显示器的大和小。如果只想知道大显示器和小显示器对工作积极性的不同影响，那么一个自变量取两个值就足够了。

值得注意的是，很多初学者往往忽视了自变量必须至少有两个取值。“使用大显示器可以提高工作积极性”这个假设实际上说的是“使用大显示器的员工工作积极性比使用小显示器的员工高”。这里显示器作为自变量，有“大”“小”两个取值。如果你的假设是“女性喜欢和人打交道的工作”，这就不能构成一个假设，因为这里的自变量只有一个取值。你可以把这个假设修正成“女性比男性更喜欢和人打交道的工作”，这就成了一个完整的假设，因为性别在这里作为自变量有两个取值。或者你可以把同样的假设改为“女性喜欢和人打交道的工作多于和机器打交道的工作”，如此工作的类型就成了自变量。

如果你想知道显示器大小和工作积极性是否存在非线性关系，你就需要多取几个值。比方说，你的假设是“显示器很小的时候，人们工作积极性很低；大一些的显示器能够提高员工的工作积极性；但是当显示器大到了一定程度，工作积极性就不再上升了”。为了检验这个假设，你需要最少取三个值，即电脑显示器很小、电脑显示器中等和电脑显示器很大。

可见，自变量的取值不是随机决定的，而是根据你的假设来确定的。在很多管理学和心理学的研究中，研究者更关心因变量会不会随自变量的升高而升高（降低），而比较少关心自变量和因变量之间的关系到底是线性函数，还是指数函数、幂函数，等等。如果是这样，

一般取两个自变量的水平就够了。以“员工之间认识时间越久，互相帮助的情况就越多”这个假设为例，在理想情况下，员工之间认识的时间是个连续的自变量，有非常多可能的取值。但是，如果你仅仅关心认识的时间会不会增加员工间互相帮助的情况，认识的时间只要有两个取值就够了。

如果一个假设有两个及以上的自变量，我们称这样的假设为多自变量假设。比方说，你有一个假设：工作年限短的员工使用大显示器比使用小显示器工作积极性高，但是工作年限长的员工使用两种显示器时工作积极性差不多。这就是一个有两个自变量的假设，一个自变量是显示器的大小，另外一个是工作年限。以此类推，你也可以把自变量增加到三个、四个，甚至更多。

从因变量的角度来看，我们也可以有不止一个的因变量。那什么时候我们需要多个因变量呢？有时加入另外一些和主要因变量相似的因变量，只是为了从另外的角度来加强实验的有效性；有时我们的理论本身就在关注自变量对两个以上的因变量的影响。

6.3　什么是好的假设？

一个好的科学研究，假设的检验固然重要，但首要的前提还是要有好的假设。在这一部分，我们将着重讨论什么样的假设才是好的假设。很多经典研究之所以经典，就是因为其假设回答了一个非常重要并且以往的研究都没能回答好的问题。自然而然，这些研究者也成了各自领域中的佼佼者。由此可见，提出一个好的假设是科学研究中最具魅力、也最具挑战的一步。

那么，到底什么样的假设才是好的假设呢？一个好的假设需要满足以下几个条件：

一个假设必须是能够证伪的（falsifiable）。理论上，一个假设应该是有可能被数据证明到底是正确还是错误的。比方说，“有志者事竟成”这个假设讲的是志向和成功的关系。如果我们不对“有志”和“成功”做出明确的定义，这就是一个没有办法证伪的假设。如果一个人没有成功，我们总是可以说他的志向还不够；如果一个人成功了，我们也总是可以说他有志向。所以，要想使这个假设成为一个可证伪的假设，我们必须对“在多大程度上有志向”算满足我们假设中的“有志”的条件有一个明确的定义。同样的道理，我们也必须对成功有一个明确的定义，否则一个人总是可以说自己成功了，而这里的关键是要看他的这个成功是不是符合我们假设里对“成功”的定义。

一个假设还必须具有理论上的重要性（theoretically important）。研究者应该能够在其他人的理论基础上，对他人的理论做出改进，或者提出以往理论没有研究过的新假设。所以要能提出好的假设，你还得知道别人做了些什么，并能站在巨人的肩上想问题。

一个假设还需要具备实际意义上的重要性（practically important）。也就是说，一个假设要有实用价值，能够回答现实生活中重要的问题，对现实生活有所启迪。有一些学术研究，耗费大量的研究经费，但是研究成果仅仅在学术上有贡献，而对人们的现实生活没有指导意

义。一个好的研究应该超越研究者所在的学术小圈子，能够直接或者间接地被应用到现实的大世界中去。

在评价一个假设是否具备实际意义上的重要性的时候，我们应该用发展的眼光看待它。如果一个研究在目前看来无法对现实生活有所贡献，但是它有可能在将来对我们的生活产生重要影响，这样的研究也是具备实际意义上的重要性的。我们所说的有实际意义上的重要性的假设，应该要么现在就能解决现实生活中的实际问题，要么具备未来解决实际问题的潜质。牛顿的三大定律就是一个很好的例子。虽然它在发现之初对当时人们的现实生活并没产生立竿见影的影响，但是对后人生活的贡献却是无法估量的。

一个假设还应该简洁（simple）。没有经验的研究者会有一个倾向，那就是在自己的假设中加入很多自变量，试图来研究这些变量之间的关系。但是随着自变量的增多，这些变量之间的关系就变得越来越复杂，最后也就越来越难以对因变量的变化做出合理的预测。比方说，有研究者想研究天气和绩效之间的关系。但同时他也意识到，性别、文化、睡眠、年龄等和绩效都有关系。如果他在他的假设里把这几个因素都加进去，假设就会变得非常复杂。此时对因变量变化的描述也会因为受到太多自变量的影响，而变得混杂不堪，从而导致其失去它在理论和实际意义上的重要性。毋庸置疑，实际状况中影响因变量的因素一定远远多于我们在假设里提及的自变量。可是，一个好的假设并不是要穷尽所有的因素，而是要分离出几个主要的因素。如果你试图把太多影响因变量的因素都包括进来，你的研究就会失去重点，也很难推广到其他的人群和情况中去。

一个好的假设还应该有繁衍性（fertile）。也就是说，从一个假设可以推演出很多具体的假设。比方说，有两个女孩子，一个叫小丽，一个叫小萍。小丽长得难看，小萍长得好看。她们现在在吵架。最为具体的假设是“小丽妒忌小萍”。这个假设就不是一个具备繁衍性的假设，因为你没办法把这个假设推演到其他的人群和情况中去。如果你在这个假设的基础上做了修改，形成了一个新的假设，如“长得难看的人常常妒忌长得好看的人”。这个假设就比前一个假设的繁衍性高一些，因为我们可以把这个假设推演到其他的人群中。如果你继续把你的假设改为“一个人在一个领域里面显弱了，就喜欢在另外一个领域里面争强”。这就是一个繁衍性更高的假设，我们不仅可以把这个假设推演到其他人群中，而且可以推演到其他很多领域中去。[①]

一个好的假设还应该是有趣的（interesting）。也就是说，一个好的假设要给读者一个惊喜。一篇文章读下来，读者通常有三种反应：第一种反应是，不看这篇文章我也知道这个结果，之所以没做这个研究是因为我觉得不值得做。比方说“睡眠不足情况下人们的绩效比在睡眠充足情况下低”之类的假设就属于这一类。

第二种反应是，不读这篇文章我不会想到事情是这样的，但是读了之后我会觉得：“我当时为什么没想到呢？”大多数的好文章都属于这一类。比方说我们前面提到的“人们喜欢把

① 该例子改编自March & Lave（1975）。

自己的偏好强加在别人身上”就是这一类研究。读了这样的文章人们会觉得眼前一亮，说：“对呀，有道理，有新意！”

第三种反应是，事实上文章里说的东西确实是正确的，如果我不读这篇文章我不会知道事情是这样的，不过读了之后我依然不能确信文章里说的东西是正确的。比方说，哥白尼提出地球是围着太阳转的。虽然现在我们知道哥白尼确实是正确的，但在当时的条件下，即使人们读懂了他的文章，也都难以信服。这种境界的研究确实为数不多，但这样的研究往往都是经典之作。

心理学中斯坦利·米尔格拉姆（Stanley Milgram）的服从实验就是一个这样的例子。米尔格拉姆教授在 20 世纪 60 年代做了一系列实验来研究人们对权威过度服从的现象。他在纽黑文市张贴广告，招募一些男性到耶鲁大学米尔格拉姆的实验室，参加一个关于“记忆和学习研究”的实验。当每个实验参与者到达实验室时，都会发现里面已经有两个人在了，一个是穿着实验室制服的实验人员，一个是叫“华莱士”（Wallace）的中年人。实际上，华莱士先生是事先安排好的，但是参加实验的人并不知情，他们以为华莱士先生是和自己一样报名参加实验的。穿着制服的实验人员向实验参与者解释，这个实验是要检验惩罚对学习效果的影响。每轮实验有两个人参加，一个人扮演“教师”的角色，另外一个人扮演“学生”。 如果“学生”回答错误的话，“教师”会对学生实施惩罚。然后实验参与者和华莱士先生抽签决定到底谁是“教师”、谁是“学生”。但实际上，这个抽签是事先做过手脚的，最后总是华莱士先生扮演“学生”，而被招募来的实验参与者总是扮演“教师”。

实验者在华莱士先生身上连上电极，并让“教师”坐在一个机器面前。这个机器上有很多按钮，不同的按钮代表不同的电压。只要按下某个电钮，华莱士先生就会被对应的电压击中——以此作为惩罚。这些按钮从 15 伏开始，最高的达 450 伏。这些按钮边上也注明有“轻微电击”“中度电击”，一直上升到“危险：严重电击”，最后超过 400 伏的按钮边是大大的红叉，以示特别警告。

“学生”华莱士先生在实验中要学习一些词组，然后回答哪些词应该是归在一组的。如果答错，“教师”就给华莱士先生一次电击。第一次电击从最低的 15 伏开始，第二次是 30 伏，之后逐渐上升。在实验中，华莱士先生实际上是从来没受到过电击的，但是“教师”并不知道。在实验中，华莱士先生会不断犯错误，受到的电击也越来越高。超过 150 伏之后，华莱士先生会发出惨叫，并要求退出实验。这个时候很多“教师”就要求停止实验。他们表示很担心华莱士先生。但是，实验者总是说：“请继续，所有的责任由我来承担。”

实际上，这个实验是来检验人们会不会服从实验者并给华莱士先生更高电压的电击。实验发现，尽管实验者只是用很简单的词句，比方说“请继续”，来要求参加实验的人继续实验，但大约有 65% 的人顺从了实验者并最终按下了高达 450 伏的按钮。实验结果大大出乎人们的意料。即使实验结果摆在那里，人们还是很难相信有高达 65% 的人对华莱士先生给出了 450 伏的电击。

一个假设要让读者产生第三种反应确实可遇不可求，但是，作为研究者，我们要尽量避免做第一种研究，争取做让读者觉得有意义并有趣的研究。

6.4　实验室研究

提出了假设之后，就要来验证它是否正确。科学发展到现在，已经有了很多检验假设的方法。我们接下来看先介绍一下在社会科学中常用的检验假设的三种方法，然后再简要介绍一下它们之间的相对利弊，最后着重介绍实验室实验的研究方法。

6.4.1　观察性研究

试想现在你有这样一个假设：同样一项活动，不付钱比付钱更能调动人们参与的积极性。那么怎样来检验这个假设呢？一个可能的方法是搜集自然发生的数据进行分析，这就是观察性研究（observational study）。比方说，在某些国家献血是无偿的，但是在另外一些国家献血是有补偿的，那么作为观察性研究，我们可以通过搜集比较这两个国家里献血的比例来检验我们的假设。

在一项新的研究开始之初，观察性研究是非常有用处的。搜集自然发生的数据可以帮助研究者对自己所要研究的问题有一个大致的了解。比方说，如果你想研究在工作中员工之间互相帮助的关系是怎样形成的，那么，首先在一些企业中对员工之间的相互帮助行为进行观察会对研究者找到最关键的因素非常有帮助。

当然，观察性研究的优越性并不仅仅局限于一项研究工作的开始阶段。如果一项研究主要在实验室里进行，那么在获得了实验室数据之后，再回到现实生活中进行实地研究可以帮助我们证实在实验室里获得的结论是否可以推广至现实环境。比方说，在实验室的环境下，你发现女性员工比男性员工更容易获得同事的帮助，那么在现实的工作环境下是否也是如此呢？实地观察性研究可以帮助我们回答这个问题。

但是，这种自然发生的数据也有它的不足。首先，自然发生的数据会受到很多和我们的假设无关的因素的影响。在“献血与补偿”的例子里，一个国家有没有献血的传统，人们对献血是不是有害健康的看法等，都会影响献血人口占总人口的比例。而由于这些因素的影响，我们就没有办法清楚地分辨出献血人口比例高低到底是由于有无补偿还是由于其他因素造成的。其次，这些自然发生的数据只能说明两个变量之间的相关关系，而不能确认两者之间的因果关系。比如，我们搜集了一组关于人们的开心程度的数据，同时也搜集了这些人朋友多少的数据。我们通过对数据的分析发现，整体来看，朋友多的人比朋友少的人更开心。但是这些数据并不能帮助我们确认，到底是因为朋友多，所以人们更加开心，还是因为人们更加开心，所以他们更容易交到更多的朋友。也就是说，通过这些自然发生的数据，我们只能说“两个变量是相关的”，但是没有办法确认变量之间的因果关系。

此外，观察性实验的结果主要取决于观察者如何理解他所观察到的现象。当被观察的因

变量是一个相对主观的变量的时候，所记录的结果会受到观察者主观解读的影响。比如，如果你的因变量是员工的高兴程度，那么观察者所记录的员工的高兴程度有很大可能与员工真实的高兴程度不相符。鉴于以上的原因，研究者通常不是通过搜集自然发生的数据，而是通过实验的方式来对假设进行检验。

6.4.2　实验室实验

正如我们前面提到的，假设描述了变量之间的因果关系。为了保证我们的实验确实能够检验自变量和因变量之间的因果关系，进行实验室实验（lab experiment）会是一个比较好的选择。相比观察性研究，在实验室实验中，我们能够更好地对其他的因素加以严格地控制，只改变我们希望改变的自变量，并监测因变量由此发生的变化。

举例来说，针对“献血与补偿”的例子，我们可以把参加实验的人聚集到实验室里面，然后把他们随机分配到有补偿和没有补偿的两个实验组中去。我们告诉有补偿组的人们，如果他们参加献血，可以得到 100 元的金钱补偿；同时我们告诉没有补偿组的人们，他们参加献血是无偿的。然后我们请这些参加实验的人回答，他们参加当前条件下的献血的可能性有多大。通常在实验室实验中，一个自变量总是取几个可能的值，而针对这些可能值的情况就是实验组。上面的实验中涉及两个实验组：一组是献血有补偿的情况，一组是献血没有补偿的情况。实验组这个概念，我们在后面的部分会经常提到。

6.4.3　实地实验

实地实验（field experiment），又称田野实验，是在自然环境下进行的有控制的实验。实验者在自然环境下控制自变量，来检验自变量的变化对因变量造成的影响，从而发现自变量和因变量之间的因果关系。本书第 8 章对于实地实验有更全面的描述。同样是检验有没有补偿对献血积极性的影响，如果是实地实验，实验者可以采用向路人发放献血宣传单的方式。宣传单有两种：一种承诺献血的人会得到金钱补偿，另外一种没有承诺金钱补偿。实验者把这两种不同的宣传单随机发给路人。然后实验者可以记录在有补偿和没有补偿的两种情况下，收到宣传单的人分别有多少人来参加献血。

有的时候，一个假设所涉及的自变量不是研究者都能控制的，比如，性别、种族、年龄等。如果我们有一个假设：男性比女性在工作中更加容易受到天气的影响。要检验这样一个假设，我们需要让一组男性和一组女性分别参加我们的实验。在这里，一个人到底是男性还是女性是不受实验者控制的，所以我们没有办法在实验中随机分配所有被试。我们把这种实验者不能直接控制自变量、不能对被试在各个实验组之间随机分配的实验叫作准实验（quasi-experiment）。本书第 7 章将对准实验做详细的介绍。

6.4.4　内部效度和外部效度

每一种研究方法都有自己的优点与缺点，不能简单地认为一种方法优于另一种方法。但

是在特定的研究需求和条件下，某种研究方法可能会比其他研究方法更适合。作为实验人员，我们需要在各个优点和缺点之间做出取舍。一方面，我们希望一个实验越接近现实越好，进而获得高的外部效度；另一方面，我们也希望能够尽可能多地对实验有更多的控制，希望提高实验的内部效度。

一个实验的内部效度是指在多大程度上我们能够确认因变量的变化确实是由自变量的变化引起的（Cook & Campbell，1979）。在一个实验中，我们关注的是自变量和因变量之间的因果关系，也就是说，我们希望能够通过实验确认因变量的变化是否是由自变量的变化引起的。如果除自变量在不同的组间发生变化之外，还有其他的因素也发生了变化，我们就没有办法确定因变量的变化确实是由自变量变化引起的。

那么，在实验室里我们如何对实验中的无关因素进行有效的控制呢？实验室实验的一大“秘诀”就是随机分配（random assignment）。随机分配指实验材料（包括被试）在各个实验组之间的分配，被试的实验顺序等是随机产生的。如果这些因素都是随机的，那么我们称之为完全随机化（complete randomization）。我们可以用电脑里的各种统计软件或者简单的随机数发生器来进行随机化操作。

做实验为什么要做到随机分配呢？随机化首先是统计分析的需要。统计分析中要求基础分析量，比如观测值（observations）和误差（errors）是独立随机变量，也就是说误差的大小独立于观测值的大小。 对被试随机分配后，我们可以认为误差是独立的、随机的，不随实验组的变化而变化，不会对因变量的值造成系统性的影响。更重要的是，随机化可以减少甚至去除某些额外因素（extraneous factors）的影响，尤其是没有得到控制的干扰因素的影响。在样本足够大时，将被试随机分到两个实验组就可以基本消除这种影响。换句话说，当样本足够大时，随机分配被试可以大大降低诸如被试的年龄等无关因素产生系统性误差的可能性。比如，在“献血与补偿”的例子中，随机分配被试可以保证被试的平均年龄在有补偿组和无补偿组都大致相同。如果不进行随机分配，就有可能存在年龄在 30 岁以上和 30 岁以下的被试被分别分到有补偿组和无补偿组中去的情况。这样，年龄作为一个额外因素就会影响实验结果。实验结果可能显示没有补偿的实验组献血更积极，但是这个结论是站不住脚的，因为更高的献血积极性可能是由年龄造成的，而不是由没有补偿造成的。

需要指出的是，随机分配必须在所有的实验组之间进行。我们再以“献血与补偿”的例子来说明。一开始你只有两个实验组：有补偿组和无补偿组。你对被试在两个实验组之间进行了随机分配。但是后来你意识到，你其实还希望了解如果补偿采取礼物而不是金钱的形式，是否会影响献血的积极性。所以，你就又找了一些被试，把他们分配到了礼物补偿组，然后比较这三个组的献血人数。但是，这样做是不对的，因为三个组的被试不是随机分配的。你必须重新做你的实验，随机在三个实验组之间分配被试。你会问：“为什么要这样自找麻烦呢？”这是因为，如果你是在做完无补偿组和金钱补偿组的实验之后，再单独加入一个礼物补偿组，那么这个组的被试有可能和你第一次做实验用的被试存在系统性差异，从而影响你

的实验结果。比方说，也许礼物补偿组的被试都是年轻人，那么这一组的被试总体就比另外两个实验组的被试年轻，你的结论自然也就不准确了。

正是因为实验室实验可以做到完全的随机分配，所以实验室实验可以达到比较高的内部效度。不过实验室实验也有明显的缺点。相对实地实验来说，实验室实验的外部效度较低。外部效度是指在多大程度上一个实验的结果能从它自身的被试和实验环境中被扩展到其他的被试和实验环境中去（Cook & Campbell，1979）。在实验室实验中，实验员营造了特殊的实验环境和条件，使被试和实验过程都处在一个“非自然态”。此外，因实验室受自身规模和经费等条件所限，测试样本难以完备，所以外部效度可能会比较低。一个实验者总是希望他得到的实验结果能够代表一个普遍的现象，而不是仅仅发生在实验参与者身上，因此我们很关心实验的可复制性（replicability），也就是你的实验结果是不是在不同的被试和实验环境下仍旧能够被重复证实。如果一个实验结果只对某一个学校的学生有效，这样的研究结果必然不具备理论意义上的重要性。

6.4.5　变量控制和测量的现实性

实验的现实性（mundane realism）由 Aronson & Carlsmith（1968）提出。它指的是实验里的情境在多大程度上也可能在被试的正常生活中发生（Aronson et al.，1998）。也就是说，高现实性的实验通常会模拟人们在日常生活中的一些经历，而不是采用人们很少会遇到的情况作为研究情境。比如，在 Asch（1951）的一个实验里，为了研究人们的判断在多大程度上会被其他人的判断所影响，被试被要求判断一组线段的长度。这些线段的长度是一目了然的。但是在他们做出判断之前，实验人员会告诉被试其他人对这组线段长度的判断，而且别人的判断是明显错误的。这个实验的现实性不是很高，因为在人们的日常生活中，很多人对一个明显的问题答案都是错误的情况并不多见。

在最近几年，组织行为学的研究越来越注重实验所涉及情境的现实性，因为高现实性的实验通常具备更高的外部效度，更容易把实验结论扩展到更广泛的情境中。在现有的组织行为学研究中，有不少研究会给被试提供一个简化的情境，然后要求被试对这个情境做出反应。这样的虚拟情境所带来的行为后果和真实情境带来的后果可能会截然不同。比如，如果我们要求被试想象一下一个应聘者长得好看，和真的有一个长得好看的应聘者坐在被试面前，是完全不同的感觉。我们要求被试想象一下天气很冷和真的在寒冷的温度下回答问题的结果也很有可能是不同的（Zhang & Risen，2016）。

组织行为学的研究也越来越关注实验采用的因变量是否涉及人的真实行为。有的因变量只能反映人们“觉得”他们会如何做决定，而不是人们“实际上”会怎样做决定。比如，消费者购买行为的研究会要求被试在一个量表上给购买可能性打分，组织行为的研究会要求被试在量表上为某个应聘者打分，等等。这样的研究经常会面临一个问题，就是无法确认人们是否会真的做出和量表选项一样的决策。比如，一个人说他“愿意献血”和他真的去献血之

间存在着巨大的不确定性，一个经理说他会雇用某个应聘者也不等于他真的会雇用这个人。

因此，组织行为学的研究应该注重在实验里采用行为变量，也就是能带来某些真实后果的行为，而不是让被试只是在问卷上填写他们会怎么做。最常见的行为变量包括能体现理论上的因变量的各种行为，比如人们的各种选择（是或否）以及涉及数量的行为（锻炼多长时间）。行为变量也可以是其他一些行为的相关变量，比如打字速度、面部表情、声音变化、荷尔蒙水平变化、眼动数据、皮肤导电程度、反应速度等。采用这样的因变量会使研究结果更加可信，毕竟我们关注的是人们真正的表现，而不是仅仅在口头上说一说或者在问卷上填一填。

当然，虽然实地实验的现实性一般都比较高，但这并不意味着我们无法在实验室实验中达到较高的现实性。在实验室实验中，我们也可以给被试设定他们熟悉的决策情境，并且观察他们的真实行为。比如，我们可以让被试选择他们想要在接下来的 10 分钟内想要听的歌曲，并根据他们的选择播放歌曲。这样的一个决策情境不仅在自变量上具备较高的现实性，而且在因变量上也涉及了人的真实行为，而且被试必须承受他们的决策的后果（播放他们选中的歌曲）。

研究者需要在一个研究的起始阶段就规划好采用什么样的行为因变量。比如，如果你的研究假设把员工对公司的喜欢程度作为一个因变量，这个假设就没有具体的行为变量。你需要在实验中展现员工对公司的喜欢程度是怎么通过他们的行为展现出来的，比如，员工会更愿意向别人推荐自己的公司，会愿意接受更远的通勤距离，等等。如果你等到研究的后期才考虑这个问题，你会发现你早期的很多实验都需要重新做一遍，因为你没有加入现实的行为变量。

6.4.6　内部效度和外部效度的权衡

如果一个实验的内部效度和外部效度都很高，自然是再好不过了。但是多数情况下内部效度和外部效度是一对矛盾体，很难在同一次实验中做到两全。在不能做到两全其美的情况下，如果一项研究更加关注两个变量之间的因果关系，那么实验室实验会是一个更好的选择，因为在实验室中我们可以通过各种手段来排除其他无关因素的影响。实际上，内部效度高是外部效度高的必要非充分条件。在必要的情况下，我们可以首先在实验室里对假设进行检验，以明确自变量与因变量的因果关系，然后在自然环境中用实地实验的方法再次进行实验，来检测这个假设的外部效度。

实地实验的外部效度通常会高于实验室实验。在实地实验中，我们通常都是使用被试的实际行为作为因变量，而不是将在实验室实验里面常常用到的“可能性”作为因变量。即使实地实验和实验室实验都使用了实际行为作为因变量，实地实验还是有它的优势：和实验室实验相比，实地实验在一个自然环境下发生，被试的决策和行为也是相对自然的。

上面我们提到，完全的随机分配是高内部效度的基石。如果一个实验能够做到完全的随

机分配，而且在实地对人们进行在自然状态下的行为的测量，那么这个实验就同时具有高的内部效度和外部效度。

同时包含实验室实验和实地实验的文章最近非常受欢迎，主要原因还是因为这样的文章兼具比较高的内部效度和比较高的外部效度。读者不仅可以确定变量之间的因果关系，而且能够确信文章里提到的现象在现实生活中确实存在，而不是在一个虚假的实验室环境下创造出来的。

当然，我们也知道，做好实地实验的挑战性是很高的。其中的一个挑战是，在实地实验中，要做到完全的随机分配比较难。比如上面提到的献血实验，研究者必须保证，看到两个不同版本的宣传单的被试之间不能交流，不然他们就会发现各自收到的宣传单不同。如果你想在一个企业里面测试两种不同的工资结构对员工绩效的影响，你也必须确定员工之间不能就工资结构进行交流。

实地实验的另外一个挑战就是很多时候研究者难以找到合适的行为因变量。一种情况是你的自变量对因变量的影响效果比较小，行为因变量难以体现出自变量的影响。比方说，如果你想研究工资结构如何影响员工对公司的忠诚度，这里，你把员工是否离职作为一个测量员工忠诚度的行为变量。我们知道，员工是不是喜欢他们的工作内容，是不是能和他们的同事愉快相处，甚至交通是否方便都会在很大程度上影响其是否离职，而工资结构只是其中的一个因素。这样，工资结构对员工是否离职的影响就有可能微乎其微，很难得到显著的统计结果。但是，这并不意味着工资结构不影响员工对公司的忠诚度。也许你使用其他的行为变量（如员工是否持续使用或购买公司产品）更有可能发现工资结构对员工忠诚度的影响。另外一种情况是在实地实验中，行为因变量受到非常多因素的影响，你需要一个非常大的样本来确保其他无关因素对因变量的影响在各实验组中是相同的。研究工资结构如何影响员工对公司的忠诚度时，你必须要有一个非常大的样本保证其他诸如工作内容、同事相处、交通等因素条件在两个实验组之间是相同的。在一个小规模的企业内部，即使你将员工随机分配到两个实验组，也很难保证其他因素完全相同。

这里，我们再讨论一个经常被忽视的问题：我们是不是永远都要追求高的外部效度呢？也许你曾经了解过一些心理学实验，其中的操作并不具备高的外部效度。比如，我们让被试记住一连串的八个数字，然后检测被试在决策中是否更加受到锚定效应的影响（Epley & Gilovich，2006）。在现实生活中，我们在做决策的时候真的会被要求必须记住一连串的八个数字吗？这种情况真的很少见，所以很明显，这样的实验外部效度不高。那么，这样的心理学实验的价值在哪里呢？

需要强调的是，很多时候，一个实验很难做到同时具有高的外部效度和内部效度。在无法两全的情况下，一项研究到底应该追求高的内部效度还是高的外部效度，要取决于研究的目的。如果一个实验的目的是要检测一种心理机制，那么这个实验并不需要有高的外部效度。实地实验通常具备非常高的外部效度，但是实地实验很少能明确地检验导致试验结果的心理

机制（Morales et al.，2017）。比方说，上面提到的实验的目的是要检验锚定效应是否是由于人们被锚定后，对自己的最终答案修正不够造成的，所以这是一个单纯的想要测试锚定效应的心理机制的实验。为了显示锚定效应的心理机制，实验者需要人工创造一个环境让被试能够把这个心理机制明显地表现出来。而在日常生活中，很多情况下这样的心理机制是被掩藏在各种其他因素里的，一个高的外部效度的实验无法把这样的心理机制完整清晰地展现出来。

6.4.7　威胁实验效度的因素

在实验中，有些实验方式或事件会影响效度，我们把这些实验方式或事件称作效度威胁因素（threats to validity）。其中有些因素会影响一个实验的内部效度，有些会影响外部效度。混淆变量（confounding variable）通常指的是没有得到控制的无关变量，这些变量使测试结果产生了系统性偏差，导致我们不能确定因变量的变化是否是由自变量的变化产生的。混淆变量是影响效度的最主要的因素。关于这部分，我们会在“如何把假设转化成实验”部分再做讨论。

下面我们讨论其他几个影响实验效度的常见因素。

被试选择偏差（subject selection bias），指被试因被主观意愿或客观条件左右，而进入不同的实验组所造成的偏差。比如，在研究工资和教育程度的相关性时，我们希望把所有样本的工资和教育程度放在一起研究。但是在现实中，当工资低于某个水平时，有些人会选择不工作。对于他们，我们可以了解他们的教育程度，却不知道如果他们工作工资会是多少。那么如果在样本中只研究有工作的人群，最后得到的工资和教育程度的相关性会与真实情况有差别，从而低估教育程度对工资的影响。所以我们利用志愿者做研究时，就要特别注意被试选择偏差问题。在研究工资与教育程度等无法避免选择偏差的情况下，有些特别的处理方法也许会有效，例如，Heckman（1974）提出的处理被试选择偏差的方法。

实验者偏差（experimenter bias），指由于实验者本身的行为所导致的偏差。比方说，如果实验操作者事先知道所要检验的假设，在进行实验的过程中，就可能有意或者无意地做出某些行为，从而影响不同实验情况下的被试反应。另外，在对一些主观数据进行编码的时候，实验编码者也可能由于知道所要检验的假设，使这些主观数据的编码存在某种倾向性。这些都会影响实验的最终结果。为了去除实验者偏差，我们通常要求不能让被试了解实验所要检验的假设。而且通常提出假设的研究者本人也不能担当实验者的角色，我们需要一个不知道所要检验的假设的人来执行实验。

成熟程度（maturation），即随着年龄的增长，被试的心理和生理会逐渐成熟，进而对实验产生影响。一般只有实验周期很长时，我们才需要考虑这种影响。当被试是儿童时，我们要特别注意这种影响。比方说，有些研究表明，即使没有接受任何治疗，大多数大学生也会在六个月内走出心理消沉期。如果有人做新药剂实验，实验结果表明服用药剂的大学生会在六个月内从心理消沉期走出来，那么我们显然不能认为药剂有疗效。我们可以采用随机化的

对照实验组来解决这个问题。

退出和减员（attrition and mortality），即在实验中，一些被试可能会退出实验，从而影响实验结果。这种情况在长期实验（longitudinal experiments）中非常普遍。如果一个实验需要被试在下个月再回实验室来回答问题，很多被试并不会按照要求回来。在组织研究中，也有被试突然被公司调去外地，不能继续参加实验的情况。因为不知道退出的被试与其他完成实验的被试有什么区别，我们很难预测这种退出和减员会对实验结果造成什么影响。最好的情况当然是尽可能消除退出和减员的情况，但是很多时候我们没有办法完全避免。这在数据分析上给我们带来很大的挑战，这时我们需要用一些统计的方法来测试退出的被试是如何影响实验结果的。

污染（contamination），指在正式实验之前进行相关度比较高的预实验（pretest）可能会使被试对实验更加熟悉和敏感，从而改变他们在正式实验里的表现。所以，预实验和正式实验应该尽量邀请不同的被试。

中值回归（regression to the mean），其典型情况是研究极端组时，测试值的变化会比研究一般群体时大得多。属于极端组的被试在下一次测试中很可能会向均值靠近。比方说，某一次测试中分数在 95 分以上的群体（满分 100），再重新接受测试时他们的分数就非常可能向均值靠近一些，平均分数通常达不到 95 分。这是一种统计学的现象，被试的分数在第二次测试中更靠近均值并不意味着这是由任何心理机制导致的。

样本不具代表性（non-representative sample），指作为样本（sample）的被试不能代表母体（population）。比如研究中国电视广告对消费者购物倾向的影响时，如果只研究汽车类广告对消费者购物倾向的影响，这样的样本就不具有代表性。实际上，很多其他因素都可以使样本不具有代表性。保证样本具有代表性是保证外部效度的基石。

霍桑效应（Hawthorne effect），指当研究人员在场时，由于紧张等原因，被试的表现会与平时不一样，这自然会影响结果的外部效度。如果我们不知道这种差别是否会对测试结果产生重大影响，那么应该怎么处理呢？一个取巧的方法是再安排一个控制组，控制组与实验组一样会被观察，但是不需要接受测试，目的只是测试霍桑效应。当然，如果你的假设决定了你有两个实验组，除非有特殊理由，一般大家认为两个实验组都会受到霍桑效应的影响，而且影响的大小应该大致相同。如果你只关心这两组之间的区别，而不是每个组测试结果的绝对值，就不需要添加控制组。因为霍桑效应只会影响两组数据的绝对值，而不会影响两组数据的相对值。但是，如果霍桑效应有可能会完全掩盖你希望检验的行为，即使加入控制组，你还是可能得不到理想的预测结果。在这样的情况下，你就需要考虑如何消除霍桑效应。比方说，为了让被试感觉他们是在正常环境下做某些行为，你可以在被试看不到的地方观察他们。

需求特性（demand characteristics），指被试在参与实验时会很自然地去猜测实验者到底想要检验什么，在实验中能引导被试做出猜测的线索被称为需求特性（Schweigert，2006）。

一旦被试对假设做出猜测，他们在实验中的行为便会或多或少受到影响。一些被试会根据他们对假设的猜测故意做出和假设一致的行为，而另外一些人也许会故意做出跟他们的猜测相反的行为。比如，在“显示器与工作积极性”的研究中，被试认为实验者想检验的假设是“显示器越大，工作积极性越高”，那么即使事实上他们的工作积极性和显示器的大小没多大关联，他们还是努力表现得和实验者的假设一致，其实他们是希望公司管理层看到实验结果后给他们配置更大的显示器。再比如，如果人们猜测实验者要检验的假设是“惩罚越多，工作表现越好”，但是因为他们不想受到惩罚，所以在受到惩罚时故意降低自己的工作表现。这就属于故意做出和假设相反行为的例子。不管他们的行为和假设是一致还是相反，实验结果的效度都受到了影响。因此，为了减少需求特性的影响，实验通常需要进行缜密的设计，比较好地隐藏实验者的真实意图，避免被被试猜测出真实的实验意图并有意调整自己的行为。

安慰剂效应（placebo effect），指被试即使没有真的接受实验，也会给出有效果的反馈。最典型的例子是药剂实验。假定你告诉被试他们吃的是止痛片，但是实际上只是维生素 C。很有可能被试也觉得疼痛减轻了，很明显，那并不是因为维生素 C 可以止痛，而是因为人们认为他们吃的是止痛片，心理上就感觉不痛了。这就是安慰剂效应。那么，如果一个被试吃了真的止痛片，痛觉减轻了，这是不是说明这个止痛片起作用呢？不是的，因为痛觉的减轻也有可能是因为被试觉得他吃了药感觉更好而已。那么，如何测出止痛片的真实效果呢？你需要有一个控制组，告诉被试他们吃的是止痛片，但是实际上给他们吃维生素 C。如果实验组的数据好于控制组，你才能得出止痛药确实有效的结论。

霍桑效应、需求效应、安慰剂效应等都有一个特点，那就是被试意识到正在进行实验，所以对测试的反馈不同于未参与实验时。因此，有些人把有这个特点的因素都称为副效应（reactivity）。不让被试知道自己正在被测试自然是最好的做法。如果做不到的话，至少不能让被试知道实验目的和假设。

6.4.8　如何把假设转化成实验

在这一部分，我们首先介绍一下在实验中如何把假设变成可以操作、可以衡量的东西，然后再介绍一些实验中需要避免的问题。

首先，我们来谈一谈在实验中如何定义一个变量，以及什么叫作可操作性定义（operational definition）。一般来说，一个变量通常是一个抽象的概念，你需要把它转换成可以操作、可以衡量的形式。那么，一个实验者用来操作或衡量的关于这个变量的可以在实验中操作的形式就是可操作性定义（Cozby，2001）。有了可操作性定义，其他的研究者就可以相对容易地重复某个实验（Elmes et al.，1999）。除了可操作性好，一个好的变量定义也要能准确、有效地代表变量。比如，把电话客户服务人员的效率只定义为接电话的数量而忽视服务质量就有一定问题。

对一个变量给出可操作性定义是实验设计中相当具有挑战性的部分。变量的抽象程度不

同，确定其可操作性定义的难易程度也不同。比方说，工作时间是一个相对来说具体的变量，你只需要用工作的小时数来衡量。而工作积极性就是一个比较复杂且抽象的变量，它会涉及很多因素：员工愿意每个星期加班几个小时，员工是否愿意接受困难的任务，员工是不是能提前完成任务，等等。一个研究者可以选择工作积极性的某一个方面来作为工作积极性的可操作性定义。而给出可操作性定义的意义在于，一个研究者必须先有一个方法来有效操作或衡量这个变量，才能具体地实施一个实验。

如果你想知道情绪对工作效率的影响，那么首先你就要知道，在一个实验中，你需要怎样做来产生你需要的情绪，所谓的工作效率应该怎样来衡量。比方说，如果你的假设是相较于快乐的情绪，伤心的情绪会使工作效率降低。你如何在实验中让人们有伤心或快乐的情绪呢？也许你会说这很简单，被试来了，问问他们高兴不高兴就行了，然后高兴的人去快乐情绪组，不高兴的人去伤心情绪组。但这是错误的。因为这样的话你就不是随机分配被试了，而是根据他们的情绪分配了。这样造成的一个结果是快乐情绪组和伤心情绪组存在其他特性上的区别（如快乐情绪组的人的受教育程度高于伤心情绪组的人），我们也就没有办法确认是否是情绪的区别导致了工作效率的区别。

一个可能的办法是，你把被试随机分成两组，让这两组人来回忆他们过去的经历，一组人回忆快乐的经历，另外一组人回忆伤心的经历。这样你就有办法使参加实验的人随机产生两种不同的情绪。然后你让被试来做某种工作，比方说让他们数零件，然后看他们在规定时间内可以完成多少。回忆过去的经历和数零件就是对情绪和工作效率的一个可操作性定义。一般来说，如果我们不能根据一个假设给出相应变量的可操作性定义，那么这个假设就是没有办法证伪的。

对一个变量给出可操作性定义时必须谨慎考虑这个定义的概念有效性（construct validity）。所谓的概念有效性，指的是变量的可操作性定义是否准确地代表了你想要操作或衡量的变量。这是一个好的实验设计的核心。

从自变量的角度来说，可操作性定义通常存在以下三个问题。我们拿通过让人们回忆过去经历的办法来产生伤心或快乐的情绪这个可操作性定义来举例子。首先，这个办法可能没有用。回忆快乐或伤心的经历可能并不能让被试在当前情况下感到快乐或伤心。那么，如果你发现工作效率在两个实验组有所不同，从而认为情绪对工作效率存在影响，这个结论就错了，因为被试的情绪在两个实验组中没有区别，工作效率的差异可能是由其他因素造成的。

其次，回忆过去经历的办法确实有效地改变了被试当前的情绪，但是被引发的情绪可能不是你希望引发的情绪。比方说，回忆伤心的经历可能没有使被试更伤心，但是使他们情绪更消极。那么，如果最后你发现这些人工作效率下降，你就没有办法得出伤心情绪降低工作效率的结论，因为更有可能是消极的情绪在起作用。当你的变量比较抽象、难以量化的时候，在确定可操作性定义时要非常小心。

最后，也是最常见且最难避免的，是你的可操作性定义不仅改变了你希望改变的因素，

也同时改变了你不希望改变的因素。也就是说，你的可操作性定义引入了混淆变量。这是一个实验最容易被论文评审人诟病的情况。比方说，回忆伤心的经历不仅让被试在当前感觉更伤心，而且让他们感觉更消极。那么，如果你发现这些人工作效率下降，到底是伤心引起的呢，还是消极引起的呢？我们不得而知，所以这个实验也就不能检验你的假设了。再比方说，你的假设是吃不健康的食物会让人感觉愧疚，而为了减轻愧疚感，人们会做一些对社会有益的事情，比如帮助别人。你给一组人吃油炸薯条，给另一组人吃蔬菜沙拉。然后你检验这两组人谁更会给别人提供帮助。但是，你操作自变量的方式引入了混淆变量。比如，比起吃蔬菜沙拉，吃油炸薯条更容易让人有饱腹感。会不会人们只有在吃饱时才会去帮助别人呢？再比如，吃油炸薯条更容易让人们觉得高兴。也许是人们在心情好的情况下才会去帮助别人？所以，我们并没有办法确认是否是愧疚感在背后起作用。

由于以上提到的原因，一般比较主观的可操作性定义都要有一个或多个问题来检验这个可操作性定义是否有效，是否同时影响了其他的因素。很多实验都包含一个“操作检验”（manipulation check）的步骤，检查你的操作是否有效改变了你希望改变的变量。比如，你希望通过一个游戏让被试觉得被别人接受或拒绝，然后测量他们在接下来的群体决策中是否会选择合作，那么我们就需要在操作检验中让被试回答他们觉得被别人接受或拒绝的程度。

很多研究人员会担心操作检验会影响被试接下来的行为和感受，这是很有可能的。如果我们问了被试他们是否感觉被别人拒绝，被人拒绝的感觉就会特别明显，从而影响我们的因变量。而如果我们没有问这个操作检验的问题，也许很多人虽然觉得被拒绝了，但是并不会让这种感觉影响他们是否合作的决定。此外，操作检验也有可能暴露研究者的真实意图，影响被试在接下来的表现。为了避免这种问题，很多操作检验是在预检测中进行的。所谓的预检测，就是在正式实验开始之前，招募另外一组被试，让这些被试经历你的自变量操作，然后测量你的自变量操作是否有效地改变了你希望操作的变量。而在正式实验中的被试不需要回答操作检验的问题。因为预检测的被试是不会参加正式实验的，所以我们不需要担心操作检验的问题会影响因变量。另外一个解决办法是把操作检验的问题放在因变量的问题之后。但是，这样的操作检验会被因变量的问题所污染，因而变得不是特别准确。

为了确保一个可操作性定义没有影响其他无关的变量，有些实验也会测量那些有可能会被影响的变量，检查这些变量是否受到影响。

对于因变量，最重要的就是要确保可操作性定义正确衡量了你的因变量。比方说，你的因变量是人们有多大动力达成一个目标。为了衡量这个因变量，你请被试回答“你在多大程度上希望达成你的目标”这个问题。这可能并不是一个非常好的可操作性定义。正如你问一个要减肥的人：“你希望能减掉 10 公斤体重吗？”大概他们都会说“非常想”。但是，“希望减肥”不同于“你有多大动力将减肥付诸行动”。因此，我们在选择问题时要非常小心。

很多时候，如果一个实验可以把一个主观的因变量用一个比较客观的方式来衡量，这样的实验通常会更被认可。比如，你可以用人们每天参加锻炼的时间来衡量达成减肥目标的动

力。这样客观的可操作性定义通常称为“行为测量”（behavioral measure）。好的实验经常把主观的可操作性定义和行为测量结合起来。行为测量不仅可以证明一个理论的实际应用，而且比主观的可操作性定义更容易为被试所理解。当然，很多时候要为一个变量找到一个好的行为测量方式并不容易。比方说，你的因变量是开心程度，这本身就是一个很主观的东西，很难用一个行为的方式来测量。也许你可以测量大家笑的次数，但这个方法并不能很准确地衡量人们真正的开心程度。

除了需要注意前面提出的问题，还要特别注意的是，在考虑如何使你的变量可以操作的时候，要避免天花板效应（ceiling effect）和地板效应（floor effect）。在实验中，有的时候会产生所有的数据都集中在可能范围的最高端的情况，这叫作天花板效应。比方说，你想证实更多的奖金可以产生更高的工作积极性。你找了一批人，告诉他们说，如果他们愿意数 5 分钟的零件，你就付给他们每人 20 元钱；对另外一批人，你告诉他们，如果他们愿意数 5 分钟的零件，你就付给他们每人 40 元钱。然后你让这些人回答，他们有多大可能性愿意来数零件。然后你发现不管是给他们 20 元钱还是 40 元钱，他们愿意数零件的可能性都在 95% 左右。这是不是意味着你的假设不成立呢？并不见得。因为很有可能你的结果受到了天花板效应的影响。也就是说，本来给 20 元钱大家就已经很愿意来数零件了，再多给他们钱也不可能提高他们数零件的积极性。如果是这样，你需要把 20 元钱的奖励调低，比方说调低到 5 元钱。当然，也有可能是因为这个百分制的衡量方式不能体现工作积极性的区别，那么你可以换一个方法来衡量因变量，比方说，你问参加实验的人：“如果我给你 20 元钱，你愿意数多少分钟的零件？”对另外一组人，你可以问：“如果我给你 40 元钱，你愿意数多少分钟的零件？”这样就避免了天花板效应。和天花板效应相反的是地板效应，它是指所有的数据都集中在可能范围的最底端的情况。它的处理方法也和天花板效应相似。我们在实验中要尽量避免这两种情况的发生，否则就无法断定到底是因为自变量确实对因变量没有影响，还是自变量没有设置在合适的水平，或者因变量没有得到合理的测量。

我们的初始实验设计可能并不完备，尤其在实验复杂或者变量相对抽象的情况下。所以有的时候，实验者会事先请少数被试做一些“测试性实验”（pilot study），小规模地测试一下实验，看看是不是有一些没考虑到的问题。为了更好地达到测试的目的，在测试性实验结束后，参加测试性实验的被试通常需要回答一些和实验的因变量无关但和实验设计有关的问题，比方说，“你是否觉得我们的实验介绍得很清楚而且容易理解？”“你在实验过程中是否遇到过很难理解的情形？”……实验者也会征求被试的意见，从而知道哪里需要改动。有的时候，实验者还要求被试在参加实验的过程中做即时的口头报告，这样被试的一些反应就可以帮助实验者对实验做出必要的改动，保证在整个实验正式开始之前能够把可能出现的问题最小化。

6.4.9　对实验结果的理解

做完了实验，搜集好了数据，我们就需要对数据进行分析。如果数据的分析结果和我们的假设不一致怎么办？是不是这就意味着我们的假设是错误的呢？先不要过早下结论，让我

们来看看什么情况下我们会得到和假设不一致的结果。

当然，出现这种情况，很有可能是因为我们的假设是错误的。但这并不是唯一的解释，还有一种可能是因为我们的实验设计不妥当。比方说，被试没能很好地理解你的指示，或者是被试在实验后期比较疲劳而没有认真回答你的问题，等等。

你还要考虑你的操作是不是有效。比方说，你对“快乐”和“伤心”的可操作性定义是分别让人们听一段欢快和缓慢的音乐。如果你的音乐没有达到让被试感到“快乐”或“伤心”的效果，那么你需要考虑采用其他的办法操作自变量。

另外，你也应该考虑你对因变量的衡量是否存在问题。有的时候，并不是你想要衡量的效应不存在，而是你没有采用合适的办法来衡量这个效应。此外，我们在前面提到过，在考虑变量的可操作性定义的时候，我们要注意选取适当的取值范围，避免产生天花板效应和地板效应。如果你发现可能存在的天花板效应和地板效应有可能造成两个实验组没有区别，那你就需要改进你的可操作性定义，再重新进行你的实验。

此外，在这个时候更为重要的是，你要思考一下：“我的实验里有没有混淆变量？”消除混淆变量的影响是保证你得到可靠数据的一个非常重要的前提。所以，你应该看一看：你本来应该控制的变量是不是得到了应有的控制？有没有其他可能的变量应该得到控制，但是你当时没有注意到？样本是不是保证随机分配且消除了随机差异？真正操作实验的人是不是对待每个被试都公正且没有倾向性？还有，被试不认真回答问题也会导致你得不到你预测的结果。 有些研究者会在一个实验的末尾加入几个问题来检测被试是否认真参与了实验。Oppenheimer et al.（2009）提出的办法最近被广泛使用。具体的做法是，在一个问题里，让被试做一些选择题（如“你最喜欢的体育节目是什么”），但是在问题的尾部，我们告诉被试不要选择他们最喜欢的体育节目，反而做一些其他无关的选择，比如点击一个问题的名称。这样，没有仔细阅读的人就会回答他们最喜欢的体育节目，只有仔细阅读的人才会按照要求点击问题的名称。这种方法可以帮助我们检验仔细阅读的人和没有仔细阅读的人是否存在系统性的差异。但是我们也必须注意到，这并不是一个完美的解决方案。目前，由于这个方法被广泛采用，很多网络上的被试已经对这个问题非常熟悉了，他们已经知道了这个问题的“窍门”在哪里，这样这个问题也因此不起作用了。一个研究者应该把注意力放在如何让被试认真回答问题上，而不是简单地把没有认真回答问题的人从样本中剔除。

6.4.10　实验结果的可复制性

可复制性（replicability）是指在相同的处理下，独立重复实验可以得到类似的实验结果。首先，复制可以让实验者对实验误差有一个估计。这种估计可以帮助实验者了解测试结果是否有统计意义上的不同。其次，由统计分析性质可知，相较于一次测试，多次的复制可以帮助我们更精确地估计样本均值（sample mean）。最后，统计分析需要一定的数据量才可以达到一定的置信度，对于复杂的实验设计来说尤其如此，而复制可以提供一定的数据量。一个可

以复制的实验才有较高的说服力。

需要特别指出的是，复制和重复测量不一样，重复测量只是从测量角度提高准确度，而复制则是重新测量整个实验被试从头到尾受到的影响。比如研究运动与心律的关系时，被试运动后测量心律，休息一定时间进行同样的运动后再测量就是复制，而被试运动后两个实验员分别通过左右手动脉同时测量其心律就是重复测量。

可复制性有两个层面：第一个层面通常是直接复制，也就是采用同样的实验设计，但是在不同的时间、地点，使用不同的被试来检验我们是不是能得到同样的实验结果；第二个层面是在保证概念有效性相同的条件下，实验者会采用不同的方式来操作自变量或者衡量因变量，来检验一个假设是否成立。即使在同一篇文章中，为了增加实验结果的可靠性，研究者也会采用不同的可操作性定义、在不同的人群中抽样等办法来重复验证同一个假设。

一个实验的结果是否可以复制是非常重要的。可以这么说，如果一个实验的结果不能被复制，那我们就有理由怀疑一个假设的正确性，或者一个效应是否真的存在。但是，以往人们对复制的兴趣并不高，主要是因为一个研究者并不能因为完全重复别人的实验而发表文章。最近在社会科学领域，研究者重新燃起了对复制实验的兴趣，主要是由于两个原因：第一，最近有研究发现，一些经常被大家引用的实验结果不能被复制；第二，有极少数的研究者最近被证实作假。一些期刊开始刊载一些复制实验的论文，这些论文主要集中在第一个层面的直接复制。

如果一个复制实验的结果和以往的结果相同，那么就意味着，我们成功复制了以前的实验结果。但是，我们要注意到，如果你重复了前人所做的实验，但是并没有得到前人所得到的结果，这种情况就比较复杂。如果因此认定前人的结果真的不能被复制还为时尚早。首先，你的复制有可能和原来的实验存在一些程序上的微小差别，从而导致最终结果的不同；其次，你的结果也许是由第二类错误（在统计学里，当某个效应存在却没有能正确识别）造成的。所以，我们不能简单地因为一个实验无法重现以前的实验结果就认为它是错误的。单单一次的复制失败可能并不足以证明一个效应真的不存在。

6.4.11　网上实验

由于互联网的广泛使用，研究人员可以在网上招募被试并让被试直接通过互联网回答问题。很多实验室实验的研究结果在网上实验中能够被重复（Horton et al.，2010），这证明网上实验是有一定的效度的。网上实验搜集数据通常速度快、成本低，因此成了很多研究者的一个普遍选择。

目前，有部分研究者对网上实验仍旧存有一些疑虑，主要集中在以下四点：

第一，网上实验太便宜。其实我们认为这是一大优点。目前看来，如果一个效应能够在网上被证实存在，通常这个效应也能在实验室中被证实存在。对很多实验来说，便宜并没有显著地影响实验的结果。当然，如果你的实验是研究金钱奖励的作用，那就另当别论了。

第二，参加网上实验的被试不具有代表性。诚然，参加网上实验的被试不能代表人口整体，但是这些被试绝大多数情况下都比大学生有代表性。在网上实验普及之前，很多论文的实验都将大学生作为被试。如果研究者能够接受用大学生作为主要被试来源，那么网上实验的被试也应该可以被接受。实际上，仅仅使用大学生作为被试在某种程度上限制了实验的外部效度，因为我们无法知道实验的结果是否仅仅局限在大学生群体里。有了网上实验平台之后，研究者可以很容易地找到非大学生被试来参加实验。从这个角度看，网上实验其实给研究者提供了一个提高外部效度的机会。

第三，网上实验只能设定一个假想的情境让被试回答，不能让被试做出行为反应 。这个想法其实是错误的。在设计得当的情况下，我们甚至可以让被试在镜头前给我们唱一首歌。

第四，参加网上实验的被试注意力集中程度不高。关于这一点，有实验发现网上实验和实验室实验几乎不存在差别，也有实验发现网上实验的被试的注意力集中程度确实稍低。但是，这并非一个难以解决的问题。很多时候，稍微修改一下实验说明就会有很大帮助。我们会在接下来的“被试的参与度”部分做详细解释。

综合来说，网上实验很多时候并不比实验室实验的效果差。如果设计得当，网上实验通常会达到相对高的内部效度和外部效度。那么，设计网上实验的时候，有没有什么需要注意的地方呢？在这里，我们主要讨论三个方面。

第一，被试的参与度（participants’ involvement）。研究者没有办法看到在电脑或手机屏幕后的被试到底做了些什么。这带来几个问题。首先，网上的被试很可能没有实验室的被试认真。他们也许一边和朋友聊天，一边参加你的实验，他们也可能很快地回答你的问题，或者根本就没有认真阅读实验说明。

上面提到，这个问题并非无法解决。比如，如果你在网上的实验说明简洁明了，效果就会好很多。你不能设计大段的文字，应该尽量用图片取而代之 。另外，在网上实验中，你也应该不断地尝试用不同的方法说明你想要被试做什么，不然被试很有可能注意不到你说了什么。

其次，网上实验涉及的任务也不能太过复杂。如果你让被试写出 10 个不愉快的经历，很多被试都会选择退出实验（关于被试退出的问题请参看接下来的“选择性退出”部分）。被试即使没有选择退出，他们也很可能不会认真对待实验任务，而是应付了事。比如，Finley & Penningroth（2015）做了一个关于记忆的网上实验，他们发现，和实验室被试相比，网上被试对实验说明的理解要差一些，而且这个问题随着实验任务复杂性的增加而变得更严重。

最后，一些类别的任务没有办法在网上实现。比如，我们的任务是给被试听一段音乐或看一段电影。这样的任务存在两个问题：一是我们没有办法确认被试是否听了音乐或看了电影。他们完全可以把声音关掉，或者播放影像，但同时打开另外一个页面。二是即使被试非常合作，我们也不能保证实验过程中是否存在技术问题，如音乐或者电影是否顺利播放。对于被试来说，他们没有动力去帮实验员解决技术问题。如果出现技术问题，他们很有可能就

自动进入下一步，这样的话，我们的操作就根本没有起到作用。所以，类似的任务我们必须从技术上保证被试是真正按照你的要求做的。

第二，重复被试（repeated participants）。在实验室实验中，实验员通常会要求被试出示身份证件，如学生证或身份证，以保证同一个被试不会在一个实验中多次出现。但是，在网上实验中，检查证件就很难做到。有的被试在同一个网站可能有多个账户，而实验员很难发觉。如果同一个被试在同一个实验中参与了多次，那么除了第一次实验的结果，后面几次的结果就都被污染了。很多网站通过技术手段能够去除大部分的重复被试，但这并不完美。比如，我们可以做到让来自同一个 IP 地址的被试只填写一次问卷，但是，一个被试可以轻易地获得多个 IP 地址，比如他的电脑和手机就有不同的 IP 地址，他家里的电脑和办公室的电脑也有不同的 IP 地址。

当然，存在少数几个重复被试通常不会对实验有效性造成实质性的影响。在当今的很多实验平台上，各种实验层出不穷，被试没有必要为了多挣一点钱去重复参加同一个实验，他们完全可以选择参加各种不同的实验。

另外一个相关的问题是，有的被试也许曾经参加过很多实验，如果你的实验采用的是通用的实验框架，这些被试可能很容易地猜到你想做什么，你的实验结果可能会因此大受影响。比如，最后通牒博弈（ultimatum game）是实验经济学中一个非常经典的实验框架，很多被试可能都做过这个博弈实验。

不过，我们也必须注意到，在某些特殊情况下，重复被试的问题对实验结果会产生很严重的影响，而且这种重复并不是被试主动重复参与导致的。比如，你想研究某种广告是如何影响淘宝买家的购买行为的。我们知道，很多淘宝买家有时使用手机平台，有时使用电脑平台。如果我们仅仅能够在技术上保证同一 IP 地址的人只能看到一个版本的广告，那么同一个人在手机和电脑间切换的时候就有可能看到不同版本的广告。更为麻烦的是，这个人的购买行为很有可能发生在他看到两个不同版本的广告之后，所以你根本没有办法分辨他的购买行为是由于他看了哪个广告引起的。在这样的情况下，实验员必须在技术上保证同一个用户名只能看到一个版本的广告（淘宝的用户名不会因为手机平台或电脑平台而改变），而不是同一个 IP 地址的人看到一个版本的广告。

第三，选择性退出（selective attrition）。被试选择性退出某个实验组会对实验结果带来非常严重的影响。在实验室实验里面，这样的情况比较少见，因为除非一些特殊情况，已经来到实验室的人通常不会中途退出。

但是选择性退出在实地实验及网上实验中都比较普遍。比如，在上面提到的“献血与补偿”实验中，控制组的部分被试由于交通问题没能来到献血现场，假定这组人的献血积极性都不高，那么最后造成的结果是，控制组的整体献血比例上升，而有金钱奖励的实验组的结果没有受到影响。如果实验者没有考虑到选择性退出的问题，就会得出给金钱奖励不如不给的结论，但是这个结论很有可能是因为控制组的部分被试选择性退出造成的。

网上实验也经常遇到选择性退出的问题。最常见的情况就是，被分配到更长、更难的任务组的被试比别的被试更容易退出。这个问题最近得到了研究者的重视。比如，Zhou & Fishbach（2016）指出，当被试出于不同的原因从不同的实验组中退出的时候，实验就可能混入混淆变量，从而影响实验的内部效度，并且导致研究者得出错误的结论。为了清楚地显示出选择性退出可能造成的问题，Zhou & Fishbach（2016）做了一个非常有意思的实验。想象一下，按照常理，使用眼线笔或使用剃须泡沫，是不可能对被试的体重造成任何影响的。但是，如果使用眼线笔使很多男性感觉这个任务很奇怪从而退出实验，同时使用剃须泡沫使很多女性退出实验，那么就会造成在眼线笔的实验组有太多女性，在剃须泡沫的实验组有太多男性，而我们知道男性的体重通常大于女性，从而造成被试体重在两个实验组之间存在显著差异。为了证明这一点，他们在 Amazon Mechanical Turk（MTurk）上招募了 100 个被试。在 MTurk 上，如果有被试中途退出，系统会自动重新招募被试，直到获得 100 个被试为止。在他们的实验中，一共有 144 个 MTurk 的被试开始了实验，但是有 41 人中途退出。其中，眼线笔实验组有 32.4%（74 人中的 24 人）的被试退出，剃须泡沫实验组有 24.3%（70 人中的 17 人）的被试退出。那么，这两组被试的体重到底有没有差异呢？实验结果发现，在眼线笔实验组，被试的平均体重是 159.64 磅；而在剃须泡沫实验组，被试的平均体重是 182.08 磅，两组的体重存在显著性差异。我们知道，单单靠想象使用眼线笔或剃须泡沫是绝对不可能影响一个人的体重的，那么这两组在体重上的显著差异只能说明我们没有在两组之间对被试进行随机分配。正如 Zhou & Fishbach（2016）所预测的那样，眼线笔实验组有 42% 的女性被试，而剃须泡沫实验组的女性仅仅有 30%。这说明，眼线笔实验组有更多的男性退出了实验，而剃须泡沫实验组有更多的女性退出了实验。如果我们没有注意到这个问题，我们就会认为仅仅想象使用眼线笔或剃须泡沫就能改变一个人的体重，从而得出错误的结论。

在这个实验里，我们通过检查被试的性别，发现选择性退出会让我们做出错误的结论。但是，我们也必须意识到，很多时候，选择性退出并不一定会在性别、年龄、种族等人口统计数据上显示出来。所以，检查人口统计数据并不一定总能帮助我们发现选择性退出造成的问题。最为稳妥的办法就是想办法在实验中消除选择性退出的问题。

6.5　实验设计

在接下来的这一部分中，我们要着重讲讲怎样设计一个实验来对假设进行检验。设计一个实验首先要考虑的就是如何把被试分配到有不同自变量取值的实验组中。你可以有两种分配方式：① 把不同的被试分配到不同的自变量取值上；② 让每个被试接受所有的自变量取值。实验的设计在很大程度上取决于你的假设——你的假设有几个自变量及每个自变量各有几个取值。如果你的假设只有一个自变量，那你的实验就是最简单的组间设计（between-subjects design）或者是组内设计（within-subjects design）。如果你有两个及以上的自变量，那

么你的实验应该是因素设计（factorial design），当然，一个因素设计既可以是组间设计，也可以是组内设计，还可以是组间组内混合的设计。下面我们对这三种设计一一加以介绍。

6.5.1　组间设计

所谓组间设计，是说不同实验组的被试是不同的，即上面的第一种分配方式。假定你有这样一个假设：对于某件东西，一个人拥有之后卖出它时索要的价格要高于他拥有之前愿意支付的价格。那么，你就可以设计这样一个实验：把被试随机分成两组，你给其中一组的人每人一个杯子，另外一组人不给杯子。你请已经有杯子的人回答，如果要把这个杯子卖掉，买方至少要出多少钱他们才愿意卖；你也请没有杯子的人回答，如果要买这样的一个杯子，他们最多愿意出多少钱。这样的一个实验采用的就是第一种分配被试的方式，是一个典型的组间设计的实验。

再比如，你的假设是，正面反馈比负面反馈更能提高员工的工作绩效。那么，你可以随机分配一组人，给他们提供正面的反馈，给另外一组人负面反馈，然后你看看这两组人的工作绩效到底哪个高。和上面的例子一样，如果一组人收到了正面反馈，那么他们就不会收到负面反馈；而收到负面反馈的那组人也不可能收到正面反馈。也就是说，每个人都只能参加一个实验组，这样的设计属于组间设计。

我们前面讲过的“显示器与工作积极性”的实验也是一个组间设计的例子。一组员工使用大显示器，另外一组员工使用小显示器，我们分别测量他们的工作积极性。如果我们的自变量有多于两个的取值，那么我们就有多于两个的实验组。比方说，我们的假设是：使用大显示器可以提高工作积极性，但是显示器大到一定程度，再增大显示器就对工作积极性没有影响了。因此，我们可以有三个实验组：第一组人使用 14 英寸显示器，第二组人使用 19 英寸显示器，第三组人使用 25 英寸显示器。然后我们分别检验各组人的工作积极性。很显然，不同实验组的人使用大小不同的显示器，这也是一个组间设计，不同之处是这个实验有更多的实验组而已。

由于不同实验组中的被试之间存在个体差异，我们在分组时需要尽可能做到对被试进行随机分配，以消除差异。

6.5.2　组内设计

另外一个减少组间差异的方法就是我们上面提到的第二种分配被试的实验设计方法——组内设计。所谓组内设计，就是被试要接受所有的自变量取值。对于组内设计来说，所有的被试参加所有的实验组，被试之间的个体差异都发生在实验组之内，所以并不需要随机分配。

我们仍旧来看“显示器与工作积极性”这个例子。你可以给所有人提供小的显示器，测量他们的工作积极性；过一段时间之后，你把所有人的显示器换成大一些的显示器，再测量他们的工作积极性；然后你比较这两种情况下人们的工作积极性。由于每个人都使用过两种

显示器，这个实验设计就是一个组内设计。

一种比较常见的组内设计是测试前—测试后设计（pretest-posttest design）。比方说，你的假设是“喝酒精饮料会降低人们的反应速度”。你可以首先测试一下被试喝酒精饮料之前的反应速度，然后你让这些人喝酒精饮料，之后再让这些人做同样的测试，记录他们的反应速度。这就是一个测试前—测试后设计。同样一组人被同样的测试方法测试了两次，一次是在自变量没有被改变之前（喝酒精饮料之前），一次是在自变量被改变之后（喝酒精饮料以后）。

6.5.3　组内设计和组间设计的选择

在资源充沛的情况下，很多实验者都偏向采用组间设计。组间设计是一种比较保守的设计，因为在组间设计中不会出现一个实验组污染另外一个实验组的情况。一般来说，组内设计存在一个问题，即更可能受到需求特性的影响。很容易想象，如果一个被试回答了两个实验组的问题，他就可以相对容易地把这两个问题进行比较，也就更可能猜测出实验者的意图，从而调整自己的行为。这就影响了实验结果的真实性。比方说，你想采用组内设计的方法来检验喝酒精饮料对反应速度的影响。由于被试在喝酒精饮料之前和喝精饮料之后做的测试相同，所以他们很容易猜测出你是想检验喝酒精饮料对他们反应速度的影响。不管他们把自己的反应速度调慢还是调快，实验的结果都存在一些偏差。如果是组间设计，需求特性的影响就相对小一些。当然，在组内设计中，我们可以通过一些实验设计的技巧来减少需求特性的影响。比方说，我们可以让被试在喝酒精饮料前和喝酒精饮料后做不同的测试，比方说，都是做数学题，但是题目不同。这样被试就很难分辨实验者的真实意图，也很难分辨哪些问题是实验者真正关心的。但是，尽管我们可以减少需求特性在组内设计中的影响，组间设计仍旧是减少需求特性更简便、更可靠的实验设计方式。

组内设计的另外一个问题就是可能产生传递效应（carryover effect）。比方说，你要测试正面反馈和负面反馈对工作绩效的影响。如果采用组内设计，被试先接受正面反馈，然后我们测量他们的工作绩效；之后被试再接受负面反馈，我们再次测量他们的工作绩效。由于对因变量的测量都是通过让被试参加相同的测试，因此被试在第二次参加这个测试时的成绩会提高，但是这不一定是由反馈对绩效的影响导致的，而很有可能是由于人们在第一次参加测试时获得的一些经验被用在第二次测试中，从而提高了成绩。我们把这种传递效应也叫作练习效应（practice effect）。但是如果被试因为重复已经做过的测试而感到无聊并逐渐对测试敷衍了事的话，成绩会降低。这也不是由反馈对绩效的影响导致的，而是另一种传递效应，即疲劳效应（fatigue effect）。我们在实验中应该尽量避免练习效应和疲劳效应。避免传递效应有一些常见的方法，如让被试回答不同的测量因变量的问题。假如你想测试被试在不同环境下的记忆力，那么不要让被试背诵相同的东西，而是背诵类似但不同的东西。

如果让所有的被试都以同样的顺序经历所有实验组的实验，他们就会很容易产生传

递效应。为了减少这种情况对实验结果的影响，我们可以用 ABBA 互相抵消（ABBA counterbalancing）的方法设计实验。仍旧以反馈和绩效的关系这一假设为例。你可以对每个被试都采用这样的实验顺序：正面反馈→负面反馈→负面反馈→正面反馈（ABBA）。把正面反馈放在第一个和第四个位置可以在某种程度上避免练习效应。但是，如果你的自变量有三个取值，上面这种 ABBA 互相抵消的方法就不太可行，因为这三个取值的顺序组合有 6 种，那么被试就要经历 3（3 个可能值）×6（6 种可能顺序组合）=18 个实验组，实在是太长了！

在这种情况下，我们有没有其他办法呢？我们可以随机把被试分配到不同的实验顺序中去，我们把这种方法叫作抵消平衡法。如果是有两个取值的自变量，这两个取值的顺序排列只有两种情况：AB 和 BA。那么你可以随机选取一半被试采用 AB 的顺序，另外一半采用 BA 的顺序。比方说，有一半的人先接到正面反馈，另一半的人先接到负面反馈。需要注意的是，在抵消平衡法中，顺序是一个组间变量。如果是有三个可能值的自变量，你就要把所有的被试随机分成 6 组，每组采用一种排列顺序。不难看出，ABBA 互相抵消的方法一般来说只适用于自变量有两个取值的情况，但是抵消平衡法却适用于自变量有两个及以上取值的情况。

可是这样还是会有问题，随着自变量的取值增多，可能的顺序也在增多。比方说，3 个自变量的取值有 6 种顺序，4 个自变量的取值有 24 种顺序，5 个自变量的取值甚至有 120 种顺序！有的时候，不同自变量取值的排列顺序的数目甚至比被试人数还多，那么随机分配被试到不同的实验顺序中去的方法也就不适用了。

这个时候，我们就没有办法做到完全的平衡抵消了。我们需要采用的是一种不完全的平衡抵消法，但是我们要保证每个取值出现的次数相同，而且这些取值可能出现的位置的次数也相同。比方说，如果我们有 A、B 和 C 三个自变量的取值，那么我们要保证三个取值出现在第一位、第二位和第三位的次数相等。这种不完全平衡抵消的方法也叫拉丁方设计（Latin-square design）。

表 6–1 列出了有四个实验组（A、B、C、D）的拉丁方设计。

表6–1　有四个实验组的拉丁方设计

		顺　序			
		第一位	**第二位**	**第三位**	**第四位**
被试编号	1	A	B	C	D
	2	B	C	D	A
	3	C	D	A	B
	4	D	A	B	C

按照表 6–1 的情况，参加实验的人数需要是 4 的倍数。比方说，如果我们有 12 个被试，被试 1、被试 5、被试 9 都采用第一个实验顺序，被试 2、被试 6、被试 10 采用第二个实验顺序，以此类推。鉴于这种分配方法的复杂性，在此我们不做深入讲述，建议感兴趣的读者参考其他相关书籍。比如，罗杰·E. 科克（Roger E. Kirk）编写的 *Experimental Design：Procedures for the Behavioral Science* 一书中对拉丁方设计有详细的介绍。

此外，值得一提的是，有时由于条件限制，可能无论是抵消平衡法还是拉丁方设计法都不能使用，因此无法在实验中加以排除或控制影响实验结果的因素。在这种情况下，只有做完实验后采用协方差分析（analysis of covariance）或偏相关等方法，把影响结果的因素分析出来，以达到对额外变量的控制。这种事后用统计技术来达到控制额外变量的方法，称为统计控制（statistical control）。

此外，还存在有多个因变量的情况，或者因变量以多个问题进行测量的情况。比方说，你的假设是“伤心情绪比快乐情绪更会降低工作效率”。在测量工作效率的时候，你用到两个测试：一个是打字测试，测量被试打字的速度；另一个是挑错字测试，测量被试挑出错字的比率。这两个测试都是用来衡量工作效率的。一般情况下，实验者会让所有的被试做这两个测试。这时实验者也会面临传递效应。为了消除传递效应，也可以采用上面提到的抵消平衡法。注意，这个时候的抵消平衡法需要在每个实验组之内使用。也就是说，在快乐情绪的实验组，要有一半的人先做打字测试，再做挑错字测试，另外一半的人测试的顺序反过来。在伤心情绪的实验组也要如此。不可以在一个实验组用一个测试顺序，在另外一个实验组用另外一个测试顺序。

当然，如果组间设计完全优于组内设计的话，我们就没有必要讨论组内设计了。组内设计有它自身的优点，主要是由于不存在被试的组间差异，组内设计更容易做出显著的效果。如果我们能很好地控制其他对组内设计的不利因素，组内设计也不失为一个好的选择。

6.5.4　因素设计

以上我们介绍了只涉及一个自变量的最基本的组内设计和组间设计。一些比较复杂的设计常常涉及多于一个自变量的情况。我们把在一个实验中同时操纵两个及以上自变量的实验设计叫作因素设计。

假定你想研究“是否拥有相似背景”这一因素如何影响人们对他人行为的理解。比方说，你有这样一个假设：如果一个人表现出好的行为，那么和他有相似背景的人倾向于认为他的表现是出于其主观意图，而没有相似背景的人则不这样认为；相反，如果一个人表现出差的行为，和他有相似背景的人更倾向于认为他表现出的行为不是出于其主观意图，而没有相似背景的人则更容易认为他的行为是出于其主观意图。这个假设有两个自变量：一是是否拥有相似的背景，二是被评价的行为的好坏。这个假设的因变量是评价人认为被评价人的行为在多大程度上是出于他的主观意图。

因此，在这样一个因素设计中，我们可以同时检验多个假设，既可以看是否拥有相似的背景如何影响人们对他人行为的理解，也可以看他人行为的好坏如何影响人们对这些行为的理解。在检验是否拥有相似背景的影响的时候，我们忽略了行为好坏在这里面的影响；同样，在检验行为好坏的影响的时候，我们也忽略了是否拥有相似背景的影响。这样的分析得出来的效应叫作主要效应（main effect）。

而如果我们把两个自变量同时考虑进来，看它们之间的组合对因变量的影响，这样的分析得出的效应叫作交互效应（interaction effect）。之所以采用因素设计，是因为我们预测实验的结果会产生一个交互效应。当然，如果你关注的不是交互效应，就不需要采用因素设计，采用最简单的单变量实验设计就可以了。

根据上面的例子，我们预测有这样一个交互效应：在理解人们的好的行为的时候，和行为人有相似背景的人比没有相似背景的人更容易认为行为人的行为是出于他的主观意图；但是在理解人们的坏的行为的时候，有相似背景的人比没有相似背景的人更容易相信行为人的行为不是出于他的主观意图。那如何来进行这个实验呢？我们可以首先把所有的被试随机分成四组：

（1）有相似背景的人理解他人的好的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，但曾经和被试一同参加新员工培训。这个同事上班从来不迟到。

（2）没有相似背景的人理解他人的好的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，而且被试和这个同事在进入公司的时候在公司的不同分部接受了新员工培训。这个同事上班从来不迟到。

（3）有相似背景的人理解他人的差的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，但曾经和被试一同参加新员工培训。这个同事上个月上班迟到五次。

（4）没有相似背景的人理解他人的差的行为：我们让被试想象，他们有一个同事，被试和这个同事并不相识，而且被试和这个同事在进入公司的时候在公司的不同分部接受了新员工培训。这个同事上个月上班迟到五次。

然后我们让第一组被试和第二组被试回答这样一个问题：“你认为你的同事上班从来不迟到在多大程度上是由于他对自己有较高要求？”被试在一个 1—11 的量表上打分，11 代表“完全由于他对自己有较高要求”，1 代表“完全不是因为对自己有较高要求”。第三组被试和第四组被试回答的问题是：“你认为你的同事上个月上班迟到在多大程度上是由于他对自己没有较高要求？”类似的，被试也在一个 1—11 的量表上打分，11 代表“完全由于他对自己没有较高要求”，1 代表“完全不是因为他对自己没有较高要求”。

假定我们的实验得到了表 6–2 所示的结果：

表6–2　实验结果举例

		有无相似背景		
		有相似背景	没有相似背景	边际平均值
行为	不迟到	9	5	7
	迟到	3	6	4.5
	边际平均值	6	5.5	/

我们对每一行或者每一列求平均值，就是表 6–2 中的边际平均值（marginal average）。边际平均值是忽略一个自变量，仅仅对因变量在另外一个自变量的某一个可能值下求得的平均值。比方说，边际平均值“7”就意味着在两次对如何理解他人行为的测量中，人们认为主观意图对不迟到这个行为的影响程度是 7。我们看到，对于迟到的行为，人们认为主观意图在这里的影响程度是 4.5。类似的，我们还计算出，不论是否迟到，有相似背景的人认为主观意图对他人的行为的影响程度是 6，没有相似背景的人认为主观意图对他人的行为的影响程度是 5.5。

图 6–1 根据上面的数据画出。横轴代表有无相似背景这一自变量，纵轴代表行为在多大程度上出于主观意图这一因变量，而另外一个自变量“行为好坏”用不同样式的线段来表示。

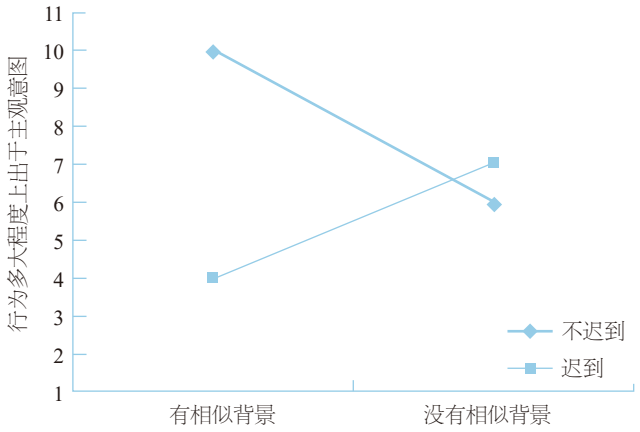

图6–1　数据的二维图——以有无相似背景为横轴

由于我们有两个自变量，但是只有一个横轴，因此我们必须决定用哪个自变量做横轴。一般来说，这取决于假设表述的形式。在上面的例子里，我们首先是固定行为的好坏，改变背景这个自变量，所以我们就把背景这个自变量作为横轴。图 6–1 显示，对于一个人的好的行为，与他有相似背景的人比没有相似背景的人更倾向于认为那是出于他的主观意图；而对于一个人的坏的行为，与他有相似背景的人比没有相似背景的人更倾向于认为那不是出于他的主观意图。

但是如果把行为好坏作为横轴，我们就得到了如图 6–2 所示的图形。

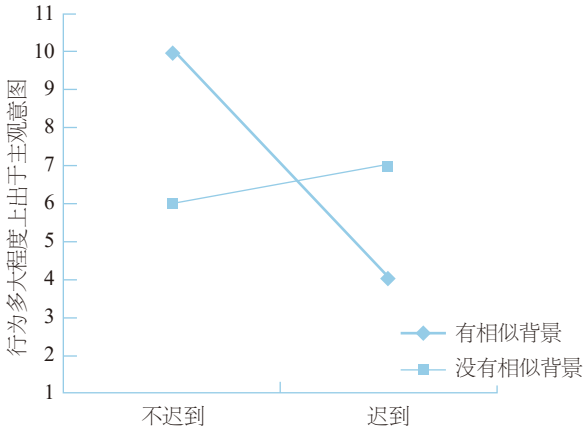

图6–2　数据的二维图——以是否迟到为横轴

对于图 6–2，比较容易的理解方式是：对于有相似背景的人，人们容易认为他人好的行为是出于其主观意图，而差的行为则不是出于其主观意图；对于没有相似背景的人，行为的好坏对人们对于他人主观意图的推测影响不大。

很多时候，图可以让人更直观地观察到变量间是否存在交互效应。一般来说，如果两条线是平行的，可以推测变量间没有交互效应；如果两条线的斜率存在较大差异，可以推测变量间是存在交互效应的。我们在下面画出了几种可能的情况。需要指出的是，交互效应的存在并不以主要效应的存在为前提。如图 6–3 所示，虽然第一张图显示的结果并不存在主要效应，也就是说，这张图的边际平均值相同，但是由于两条线的斜率明显不同，就证明了变量间存在交互效应。

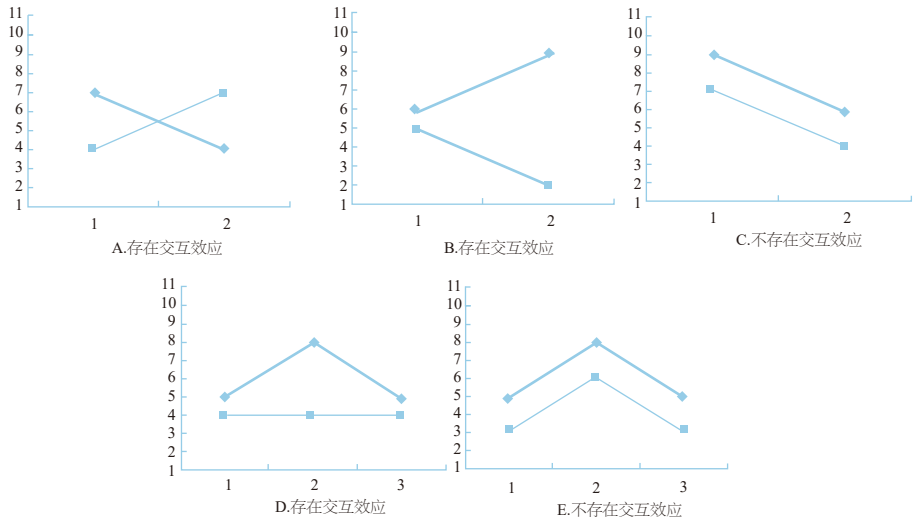

图6–3　几种可能情况图示

检验是否存在交互效应的常用方法是方差分析（analysis of variance，ANOVA），也称变异量分析。此外，如果你的实验包含了操作检验的问题，通常你也需要用 ANOVA 分析一下

一个变量操作检验的结果是不是独立于另外一个变量。也就是说，我们需要保证两个自变量是互相独立的。这时，一个变量的操作检验结果应该只受该变量的影响，而不会和另外一个变量产生交互效应。

我们上面讲到的例子是一个典型的组间因素设计（between-subjects factorial design）。必须明确的是，因素设计和组间设计、组内设计之间不是互相排斥的，一个因素设计可以是单纯的组间因素设计，也可以是组内因素设计，甚至还可以是组间组内混合的因素设计。

我们先来说说组间组内混合的因素设计。假定我们现在有两组被试：一组被试先想象一个跟他同时参加新员工培训的同事，这个同事上班从来不迟到，并让被试回答他认为这个同事上班不迟到在多大程度上是因为对自己有较高要求；然后再让被试想象一个没有跟他一起参加过新员工培训的同事，这个同事上班也从来不迟到，并让被试回答他认为这个同事上班不迟到在多大程度上是因为对自己有较高要求。另外一组被试也回答两次问题，只不过这组被试需要想象一个同事上个月上班迟到了五次。这就是一个组间组内混合的因素设计。其中，是否有相似背景这个自变量是一个组间变量，而同事的行为好坏这个变量是一个组内变量，同样的被试分别想象了两个同事，并两次回答了相同的关于因变量的问题。

如果更进一步，让被试把所有的实验组都经历一遍，那就是一个完全的组内因素设计。

到底选取组间因素设计、组内因素设计，还是混合因素设计，不是实验者可以任意决定的。它取决于你的假设和实验条件。在上面的例子中，很明显完全组内因素设计不是一个好的选择。它不仅容易产生传递效应，而且被试非常容易猜测出实验者的意图。如果我们想要保证各个实验组互不影响，减少混淆变量的影响的话，采用组间设计比较妥当。

当然，组内因素设计或者混合因素设计也有它们自身的好处。比方说，有的时候一个假设本身关注的就是组内因素的变化，这个时候就应该采用组内因素设计或者混合因素设计。比如，你想检验人们不同时间点上的心情变化，以及是否吃早饭对心情的影响。你想知道人们是不是下午比早上心情好，而且你想研究吃不吃早饭和时间（上午和下午）对人们的心情是否有交互作用。由于本身就是想比较同一个自变量在同一组人身上的变化，时间变量最好作为一个组内变量。是否吃早饭当然是作为一个组间变量，因为你不可能让人既吃早饭又不吃早饭。

如果一个因素设计有两个自变量，相对应的交互作用就叫作两重交互作用（two-way interaction）。如果我们有多于两个的自变量，这样的设计叫作高阶设计（higher-order design）。比方说，在背景和行为之间的交互影响的例子中，再加入时间这个自变量，分别在早上和晚上测量人们如何理解他人的行为，我们就有了一个 $2 \times 2 \times 2$ 的高阶设计，一共有 8 个实验组。在一个有三个自变量的设计中，假设三个自变量分别为 A、B、C。那么，这个实验会产生三个主要效应，分别对应 A、B、C。还有三个两重交互作用，分别发生在 A、B 之间，A、C 之间，以及 B、C 之间。还有一个三重交互作用，发生在 A、B、C 三个自变量之间。对于一个高阶设计来说，我们的假设关注的应该是多重交互作用，否则不必也不应该采用高阶的设计。

需要注意的是，到底是几重交互作用，取决于你有多少个自变量，而不取决于自变量值的个数。比方说，图 6–3 中的 D 图和 E 图就是一个两重交互的例子，因为这个图上只有两个自变量。虽然其中一个自变量有三个水平，但仍旧是一个两重交互作用，而不是一个三重交互作用。我们上面讲的都是每个自变量有两个取值的情况，实际上很多时候自变量有多于两个的取值。这个时候，只要我们只有两个自变量，交互作用就仍旧是两重交互作用，尽管你需要更多的实验组 。总之，实验设计中可以有多个自变量，而每个自变量又可以有多个水平，自变量既可以是组内变量，也可以是组间变量。

研究中最常见的就是两重交互作用。当自变量增多的时候，对实验结果的解释就变得困难起来。很多时候我们很难理解一个四阶的交互作用到底意味着什么。更多的自变量会混淆我们对问题的理解，而且通常不具备理论上的重要性。这时候我们可以采用一个实验设计技巧：把我们不关心的多重交互作用和区块混淆在一起（confound with block）。这种做法的指导思想是让一个区块内元素受来自同样的某种多重交互作用的影响。这样，区块的影响和多重交互作用这两种我们都不关心的，但是会影响实验结果的因素，就被放到一起考虑了，就可以把其共同作用的影响仅当作区块的影响。具体的原理和处理方法可以参考 Douglas（2005）。

6.6　结语

在本章中，我们首先介绍了研究的类型，以及什么样的假设才是一个好的假设。然后，我们着重讲述了如何在实验室中对假设进行检验。在实验室实验中，我们又着重讲了组间设计、组内设计和因素设计这三种最常见的实验设计方法和它们各自的优缺点。实验设计涉及很多概念，有许多需要注意的问题。很多好的研究不仅有好的假设，还有让人信服而且印象深刻的实验设计。实验设计本身是一门科学，同时也是一种艺术。

思考题

1. 随机分配在实验里的作用是什么？

2. 什么样的实验设计具备较高的内部效度？什么样的实验设计具备较高的外部效度？为什么有的实验的外部效度很低，但仍旧被认为是一个好的实验？

3. 为什么研究者越来越看重实验的现实性？如何最大化实验的现实性？

4. 什么是混淆变量？

5. 组间设计和组内设计各有什么优缺点？

6. 选择三篇你感兴趣的文章，思考：这些研究是怎样操作自变量的？又是怎样衡量因变量的？你觉得这些方法有改进的空间吗？为什么？

延伸阅读

Aguinis，H. & Bradley，K. J.（2014）. Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational research methods*，*17*（4），351–371.

Cook，T. D. & Campbell，D. T.（1979）. *Quasi-experimentation：Design and analysis issues for field settings.* Chicago：Rand McNally.

Haslam，S. A.，& McGarty，C.（2004）. Experimental design and causality in social psychological research. C.，Sansone，CC，Morf，AT Panter，（*Eds.*），*The Sage handbook of methods in social psychology*，237–264. Thousand Oaks：Sage.

Highhouse，S.（2009）. Designing experiments that generalize. *Organizational research methods*，*12*（3），554–566.

Kardes，F. R. & Herr，P. M.（2019）. Experimental research methods in consumer psychology. In *Handbook of research methods in consumer psychology*（pp. 3–16）. London：Routledge.

Mitchell，G.（2012）. Revisiting truth or triviality：The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*，*7*（2），109–117.

Montgomery，D. C.（2005）. *Design and Analysis of Experiments*（sixth edition）. NY：John Wiley & Sons.

Mook，D. G.（1983）. In defence of external invalidity. *American psychologist*，*38*（4），379.

Schweigert，W. A.（2006）. *Research methods in psychology*. Long Grove，IL：Waveland.

Vogt，W. P.，Gardner，D. C. & Haeffele，L. M.（2012）. *When to use what research design*. Guilford Press.

Wilson，T. D.，Aronson，E. & Carlsmith，K.（2010）. The art of laboratory experimentation. *Handbook of social psychology*，*1*，51–81.

陈晓萍 .（2017）. 实验之美：简单透彻地揭示因果关系 . 管理学季刊，*2*（02），114–126.